Bounding the Fairness and Accuracy of Classifiers from Population Statistics

Sivan Sabato [1 2] and **Elad Yom-Tov** [2]

Abstract

We consider the study of a classification model whose properties are impossible to estimate using a validation set, either due to the absence of such a set or because access to the classifier, even as a black-box, is impossible. Instead, only aggregate statistics on the rate of positive predictions in each of several sub-populations are available, as well as the true rates of positive labels in each of these sub-populations. We show that these aggregate statistics can be used to lower-bound the *discrepancy* of a classifier, which is a measure that balances inaccuracy and unfairness. To this end, we define a new measure of unfairness, equal to the fraction of the population on which the classifier behaves differently, compared to its global, ideally fair behavior, as defined by the measure of *equalized odds*. We propose an efficient and practical procedure for finding the best possible lower bound on the discrepancy of the classifier, given the aggregate statistics, and demonstrate in experiments the empirical tightness of this lower bound, as well as its possible uses on various types of problems, ranging from estimating the quality of voting polls to measuring the effectiveness of patient identification from internet search queries. The code and data are available at `https://github.com/sivansabato/bfa`.

1. Introduction

Suppose that a health insurance company uses some unpublished method to decide whether a person should be classified as "at risk" for diabetes, for instance so as to offer diabetes screening. Two desirable properties of such a classifier are accuracy and fairness. For this example, consider fairness with respect to the state of residence (assuming a US-based population). Accuracy and fairness are easy to estimate using a validation set; However, the only details which are publicly available from the insurance company are aggregate statistics on the number of people identified as "at risk" in each state. In addition, the true proportions of diabetes in each state are known. Based only on these two types of aggregate statistics—the true positive rate in each state, and the predicted positive rate in each state—what can be inferred regarding the fairness and/or accuracy of the classifier? Moreover, suppose that the company publishes some additional information about the accuracy of the classifier, or about its fairness with respect to the state of residence. Can this information be used to make any inferences about the other property?

Note that in the scenario above we used "state-of-residence" as an example of an attribute that partitions the population; The same questions can be asked with respect to any other attribute for which fairness is desired, such as race, age, religion or gender (see, e.g., Chen et al., 2019). The case of multi-valued attributes is more intricate, and is the main focus of this work.

A similar question arises when a classifier is designed or learned with little or no labeled training data, or when the available training data is not representative of the target distribution. Various methodologies allow constructing a classifier in these cases, such as unsupervised learning, transfer learning, and hand-crafting rules based on domain expertise. The challenge is to then estimate the quality of the classifier without a validation set. Here too, we are interested in both the accuracy of the classifier and its fairness with respect to an attribute of interest.

In this work, we show that the aggregate statistics described above can be used to lower-bound the *discrepancy* of a classifier, a measure that balances inaccuracy and unfairness. Following Hardt et al. (2016), we say that a binary classifier is perfectly fair if it satisfies the property of *equalized odds* (also termed *disparate mistreatment*; see, e.g., Zafar et al., 2017a). This property requires that the false positive rate (FPR) and the false negative rate (FNR), conditioned on the value of the attribute, be the same for all values. It is well known that fairness and accuracy in classification do not always co-exist: in some cases, a more accurate classifier

[1]Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva, Israel [2]Microsoft Research, Herzelia, Israel.. Correspondence to: Sivan Sabato <sabatos@cs.bgu.ac.il>, Elad Yom-Tov <eladyt@microsoft.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

implies that it must be less fair, and vice versa (Pleiss et al., 2017; Kleinberg et al., 2017; Menon & Williamson, 2018; Goel et al., 2018). In this work, we provide a method for quantifying this trade-off for a particular classifier, based only on its aggregate statistics. We define a measure of $\texttt{unfairness}$, which is equal to the fraction of the population on which the classifier behaves differently compared to its baseline, ideally fair, behavior. Considering the possible trade-offs of fairness and accuracy given the aggregate statistics, questions that can be addressed include:

- Is it possible that the classifier is fair? If not, how unfair must it be in the best possible scenario?
- Suppose that the classifier is known to be nearly fair. How accurate can it be?
- Suppose that the classifier is known to be quite accurate. How fair can it be?
- Suppose that there is a penalty for each person who gets a wrong prediction, and for each person who is treated unfairly; what is the smallest possible overall cost of this classifier?

Each of the questions above is equivalent to asking for a lower bound on

$$\texttt{discrepancy}_\beta := \beta \cdot \texttt{unfairness} + (1-\beta) \cdot \texttt{error}, \tag{1}$$

For some $\beta \in [0,1]$, or for a β that optimizes a constraint. For instance, if it is known that $\texttt{unfairness} \leq U$ for some known U, one can lower-bound $\texttt{discrepancy}_\beta$ for the smallest β such that the minimizing solution satisfies $\texttt{unfairness} \leq U$, to find the minimal value of $\texttt{error}$ under this constraint. We derive an efficient and practical procedure that finds an optimal lower bound on $\texttt{discrepancy}_\beta$ given the aggregate statistics. This procedure can help answer each of the questions above. In addition, we report experiments, which demonstrate the tightness of the lower bound and possible uses of the procedure.

Paper structure. After discussing related work, we formally define the setting and notations in Section 2. In Section 3, we define our measure of $\texttt{unfairness}$, and show how it can be efficiently calculated from known FPRs and FNRs. The main algorithmic contribution is provided in Section 4, where a practical and efficient procedure for finding an optimal lower bound for $\texttt{discrepancy}_\beta$ is derived. Experiments are reported in Section 5. We close with a discussion in Section 6.

Related work

Fairness in classification has been a highly studied topic of research in recent years, due to its importance in legal, financial, and medical decisions (Barocas et al., 2017). This importance has grown in parallel with the wide application of automated (and frequently, opaque) models in multiple areas affecting people. Various notions of fairness have been proposed (see, e.g., Dwork et al., 2012; Grgic-Hlaca et al., 2016; Kusner et al., 2017; Berk et al., 2018; Verma & Rubin, 2018). In this work, we focus on the notion of *equalized odds* (Hardt et al., 2016), which requires equal FPRs and FNRs in each sub-population, where a sub-population is the set of individuals who share the same value of the attribute of interest. Many works propose methods for learning fair classifiers under the equalized-odds definition (see, e.g., Feldman et al., 2015; Hardt et al., 2016; Goh et al., 2016; Zafar et al., 2017b;a; Woodworth et al., 2017; Wu et al., 2019). Learning methods that guarantee or approximate other definitions of fairness, such as equal opportunity and demographic fairness, have also been widely studied in recent years (e.g., Dwork et al., 2012; Zemel et al., 2013; Calmon et al., 2017b; Donini et al., 2018; Goel et al., 2018; Johndrow & Lum, 2019).

Auditing a classifier for fairness is a crucial task in the pipeline of learning fair classifiers (Bellamy et al., 2019). Given access to the classifier and its individual predictions, Black et al. (2019) propose a method for fine-grained scrutiny of a classifier beyond group fairness. McDuff et al. (2019) propose a simulation-based approach for interrogating the classifier. Kusner et al. (2017) define the property of "counterfactual fairness", which can be tested given access to the classifier or to individual classified examples.

Learning classifiers based on both individual covariates and population statistics has been studied under the title "ecological inference". Jackson et al. (2006; 2008) propose methods for regression based on both individual-level and aggregate-level statistics. Sun et al. (2017) infer voting patterns using aggregate statistics. We are not aware of works that attempt to estimate properties of existing classifiers from aggregate statistics alone, and in particular in the context of fairness.

Approaches for quantifying unfairness for individual-fairness notions have been suggested by Heidari et al. (2018); Speicher et al. (2018). For group fairness, previous works impose constraints for requiring that a classifier is almost-fair (see, e.g. Donini et al., 2018; Calmon et al., 2017a), but do not suggest an overall measure of (un)fairness.

2. Setting and Notations

We consider a binary classification problem, in which each individual in the population has a true label in $\mathcal{Y} = \{0,1\}$. In addition, we assume an attribute of interest, such as race, state of residence, or age, which assigns a value for each

individual. We denote the (finite) set of possible values of this attribute by $\mathcal{G}$. For simplicity and concreteness, we henceforth call the possible values of the attribute *regions*, alluding to location-based attributes such as state of residence or country of origin. A *sub-population* is a subset of the population which includes all the individuals in the same region.

We wish to study some existing classification model mapping each individual from the population to a label, which may or may not be equal to the true label of that individual. Denote by $\mathcal{D}$ the distribution in the population over the triplets of true label, predicted label, and region. A random triplet drawn according to $\mathcal{D}$ is denoted by $(Y, \hat{Y}, G)$, where $Y \in \mathcal{Y}$ is the true label, $\hat{Y} \in \mathcal{Y}$ is the predicted label, and $G \in \mathcal{G}$ is the region of the individual. We assume that $\mathcal{D}$ is unknown, and the only available information is in the form of aggregate statistics by region. Denote the probability of an event E by $\mathbb{P}[E]$, and the probability of the event conditioned on the region by $\mathbb{P}_g[E] := \mathbb{P}[E \mid G = g]$. The available information is the following:

- The *true positive rate* of each label $y \in \mathcal{Y}$ in each sub-population $g \in \mathcal{G}$, denoted by
$$\pi_g^y := \mathbb{P}[Y = y \mid G = g] \equiv \mathbb{P}_g[Y = y],$$
- The *predicted positive rate* of each label $y \in \mathcal{Y}$ in each sub-population $g \in \mathcal{G}$, denoted by
$$\hat{p}_g^y := \mathbb{P}[\hat{Y} = y \mid G = g] \equiv \mathbb{P}_g[\hat{Y} = y].$$
- The relative weight of each sub-population:
$$w_g := \mathbb{P}[G = g].$$

Denote the available information by

$$\texttt{Inputs} := (\{w_g\}_{g\in\mathcal{G}}, \{(\pi_g^y, \hat{p}_g^y)\}_{g\in\mathcal{G}, y\in\mathcal{Y}}).$$

Henceforth, we sometimes omit the subscripts $g \in \mathcal{G}, y \in \mathcal{Y}$ from set notations. Note that by definition, $\pi_g^1 = 1 - \pi_g^0$ and $\hat{p}_g^1 = 1 - \hat{p}_g^0$. We define the two complements as part of the input for convenience.

Denote the overall FPR and FNR of the classifier by $\alpha_{\text{all}}^0, \alpha_{\text{all}}^1$ respectively:

$$\forall y \in \mathcal{Y}, \alpha_{\text{all}}^y := \mathbb{P}[\hat{Y} \neq y \mid Y = y].$$

The FPR and the FNR of the classifier on each sub-population $g \in \mathcal{G}$ is denoted by

$$\alpha_g^y := \mathbb{P}_g[\hat{Y} \neq y \mid Y = y].$$

The population error of the classifier is given by:

$$\texttt{error} := \mathbb{P}[\hat{Y} \neq Y] = \sum_{g\in\mathcal{G}} w_g \sum_{y\in\mathcal{Y}} \pi_g^y \alpha_g^y. \qquad (2)$$

Note that if $\pi_g^y = 0$ then α_g^y is undefined, but is also not required to calculate the value of the error.

Equalized odds, in our notation, states that for each $y \in \mathcal{Y}$ and any $g, g' \in \mathcal{G}$, we have $\alpha_g^y = \alpha_{g'}^y$. This implies that for all $y \in \mathcal{Y}, g \in \mathcal{G}$, $\alpha_g^y = \alpha_{\text{all}}^y$. In many cases, however, the classifier might not be completely fair. For instance the classifier may have been constructed to approximate fairness, e.g., using one of the methods in (Goh et al., 2016; Zafar et al., 2017b). Nonetheless, being *close* to fairness as much as possible is a desired property. In the next section, we define a measure of unfairness, which quantifies the amount of unfairness of a classifier with respect to the equalized-odds fairness criterion.

3. Quantifying Unfairness

We propose a new measure of classifier unfairness, which quantifies the fraction of the population on which the classifier behaves unfairly, that is, has a different conditional distribution of predicted labels, as defined below. The unfairness of a classifier depends on its FPR and FNR in each region, given by $\{\alpha_g^y\}$. Let $y \in \mathcal{Y}$, and recall that $\mathcal{D}$ is the distribution of the triplet $(Y, \hat{Y}, G)$. Denote the conditional distribution of $\mathcal{D}$ given $G = g$ by $\mathcal{D}_g$.

For each $y \in \mathcal{Y}$, we model the conditional distribution of $\hat{Y} \mid Y = y$ under $\mathcal{D}_g$ as a mixture of two conditional distributions: a baseline distribution, which is the same for all regions g, and a local "nuisance" distribution, which can be different for each g. Let $\eta_g^y \in [0, 1]$ be the weight of the nuisance distribution for g, y, and define a random variable N which is equal to 1 if $\hat{Y}$ is drawn according to the nuisance distribution, and equal to 0 if it is drawn according to the baseline distribution. Then $\eta_g^y = \mathbb{P}_g[N = 1 \mid Y = y]$.

It is easy to see that if the classifier is fair, then for each $y \in \mathcal{Y}$, $\mathcal{D}_g$ has the same distribution conditioned on $Y = y$ for all $g \in \mathcal{G}$. In this case, $\eta_g^y = 0$ for all $g \in \mathcal{G}, y \in \mathcal{Y}$. However, if the classifier is not completely fair, then $\{\eta_g^y\}$ cannot all be zero. We define the measure of `unfairness` as the *fraction of the population that is treated differently from the baseline treatment*. This fraction is equal to $\sum_{y\in\mathcal{Y}} \sum_{g\in\mathcal{G}} w_g \pi_g^y \eta_g^y$. Since the decomposition into a baseline distribution and a nuisance distribution is unobserved, `unfairness` is defined as this value under the *best possible decomposition*. This is the decomposition which minimizes the value subject to the per-region FPR and FNR of the classifier, which are given by $\{\alpha_g^y\}_{g\in\mathcal{G}, y\in\mathcal{Y}}$. Denote the false positive rate of the baseline distribution by α^0 and the false negative rate by α^1. Formally: $\alpha^y := \mathbb{P}[\hat{Y} \neq y \mid Y = y, N = 0]$. Then the

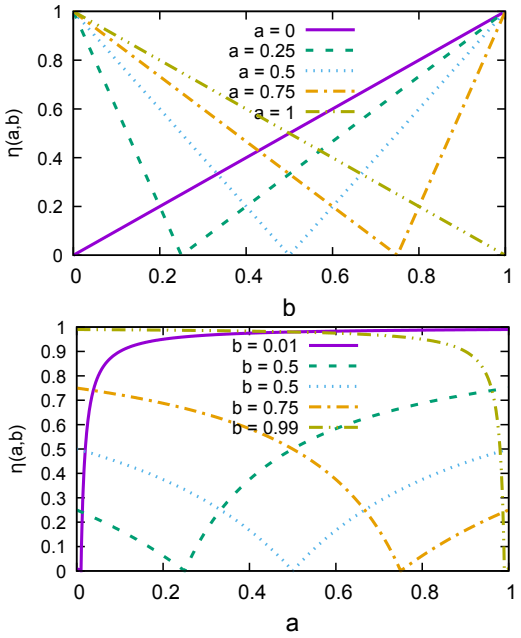


Figure 1. Left: $b \mapsto \eta(a,b)$ for various values of a. Right: $a \mapsto \eta(a,b)$ for various values of b.

following relationship holds:

$$\begin{aligned}\alpha_g^y &= \mathbb{P}_g[\hat{Y} \neq y \mid Y = y] \\ &= \quad \mathbb{P}_g[\hat{Y} \neq y \mid Y = y, N = 0] \cdot \mathbb{P}_g[N = 0 \mid Y = y] \\ &\quad + \mathbb{P}_g[\hat{Y} \neq y \mid Y = y, N = 1] \cdot \mathbb{P}_g[N = 1 \mid Y = y] \\ &= \alpha^y(1 - \eta_g^y) + \mathbb{P}_g[\hat{Y} \neq y \mid Y = y, N = 1] \cdot \eta_g^y.\end{aligned}$$

Since $\mathbb{P}_g[\hat{Y} \neq y \mid Y = y, N = 1] \in [0,1]$, we get

$$\alpha^y(1 - \eta_g^y) \leq \alpha_g^y \leq \alpha^y(1 - \eta_g^y) + \eta_g^y.$$

Thus, for $\alpha^y \in (0,1)$, we have the following lower bound: $\eta_g^y \geq \max(1 - \frac{\alpha_g^y}{\alpha^y}, 1 - \frac{1-\alpha_g^y}{1-\alpha^y})$. In addition, if $\alpha^y = 0$, then $\eta_g^y \geq \alpha_g^y$, and if $\alpha^y = 1$, then $\eta_g \geq 1 - \alpha_g^y$. Moreover, η_g^y satisfies these lower bounds with equality. This can be seen by considering a deterministic nuisance distribution which draws y if $\alpha_g^y < \alpha^y$ and $1 - y$ otherwise. Thus, given α^y and α_g^y, the minimal value of η_g^y is $\eta(\alpha^y, \alpha_g^y)$, where $\eta : [0,1]^2 \to [0,1]$ is defined as follows (see Figure 1):

$$\eta(a,b) = \begin{cases} 1 - b/a & b < a, \\ 1 - (1-b)/(1-a) & b > a, \\ 0 & b = a. \end{cases}$$

To find the value of `unfairness` for the best possible decomposition, we minimize over α^0, α^1, the FPR and FNR that determine the baseline distribution:

$$\texttt{unfairness}(\{\alpha_g^y\}) = \sum_{y \in \mathcal{Y}} \min_{\substack{\alpha^y \in \\ [0,1]^2}} \sum_{g \in \mathcal{G}} w_g \pi_g^y \eta(\alpha^y, \alpha_g^y). \quad (3)$$

We now show that this function can be easily minimized exactly. Define $\psi^y(a) := \sum_{g \in \mathcal{G}} w_g \pi_g^y \eta(a, \alpha_g^y)$. First, suppose that for all $g \in \mathcal{G}, y \in \mathcal{Y}$, we have $\alpha_g^y \in (0,1)$. For any fixed $b \in (0,1)$, $a \mapsto \eta(a,b)$ is concave on the intervals $[0,b]$ and $[b,1]$. Thus, ψ^y is concave on any closed interval which is a subset of $[0, \alpha_g^y]$ or of $[\alpha_g^y, 1]$ for all $g \in \mathcal{G}$. In each such interval, the minimizer of ψ^y is one of the end points of the interval. Therefore, ψ^y is minimized at an end point of a maximal interval which satisfies this property, that is, at a point in $A := \{\alpha_g^y\}_{g,y} \cup \{0,1\}$. Now, if for some g, y, $\alpha_g^y \in \{0,1\}$, then $\eta(a, \alpha_g^y) = \mathbf{1}_{a \neq \alpha_g^y}$. Thus, this does not add other possible minimizers to A. Hence,

$$\texttt{unfairness}(\{\alpha_g^y\}) = \sum_{y \in \mathcal{Y}} \min_{\alpha^y \in A} \sum_{g \in \mathcal{G}} w_g \pi_g^y \eta(\alpha^y, \alpha_g^y).$$

We conclude that given $\{\alpha_g^y\}$, the value of `unfairness` can be calculated exactly, in time linear in $|\mathcal{G}|$.

The `error` and the `unfairness` of a classifier, on a distribution described by $\{w_g\}$ and $\{\pi_g^y\}$, are fully determined by the values of $\{\alpha_g^y\}$ for this classifier, as can be seen in Eq. (2) and Eq. (3). However, in our setting, only the aggregate positive rates of the classifier $\{\hat{p}_g^y\}$ are provided. These do not fully determine the values of α_g^y for all $y \in \mathcal{Y}, g \in \mathcal{G}$. Nonetheless, some relationships can still be established. First, note that by definition, if $\pi_g^y = 1$ then

$$\alpha_g^y \equiv \mathbb{P}_g \hat{Y} \neq y \mid Y = y] = \mathbb{P}_g[\hat{Y} \neq y] = 1 - \hat{p}_g^y. \quad (4)$$

In addition, if $\pi_g^y = 0$ then α_g^y is undefined, but also has no bearing on `error` or any other property of the classifier. Thus, it suffices to solve for values of α_g^y for $y \in \mathcal{Y}$ and $g \in \mathcal{G}^+$, where $\mathcal{G}^+ := \{G \mid \forall y \in \mathcal{Y}, \pi_g^y \neq 1\}$. We assume that $\mathcal{G}^+$ is non-empty, otherwise the problem is trivial. Now, given `Inputs`, a simple linear relationship can be observed between α_g^0 and α_g^1 for $g \in \mathcal{G}^+$. We have

$$\begin{aligned}\hat{p}_g^1 &\equiv \mathbb{P}_g[\hat{Y} = 1] \\ &= \mathbb{P}_g[\hat{Y} = 1 \wedge Y = 1] + \mathbb{P}_g[\hat{Y} = 1 \wedge Y = 0] \\ &= \mathbb{P}_g[\hat{Y} = 1 \mid Y = 1] \cdot \mathbb{P}_g[Y = 1] \\ &\quad + \mathbb{P}_g[\hat{Y} = 1 \mid Y = 0] \cdot \mathbb{P}_g[Y = 0] \\ &= (1 - \alpha_g^1)\pi_g^1 + \alpha_g^0 \pi_g^0.\end{aligned}$$

Denoting $r_g := 1 - \hat{p}_g^1/\pi_g^1$ and $q_g := \pi_g^0/\pi_g^1 \equiv 1/\pi_g^1 - 1$, we get that, for any classifier (whether fair or not),

$$\forall g \in \mathcal{G}^+, \quad \alpha_g^1 = r_g + q_g \alpha_g^0. \quad (5)$$

r_g and q_g are well-defined for $g \in G^+$. Thus, the unknown variables are $\{\alpha_g^0\}_{g \in \mathcal{G}^+}$.

4. Lower-bounding Classifier Discrepancy from Population Statistics

Having defined the measure of `unfairness`, we now develop a procedure for lower-bounding the value of $\texttt{disc}_\beta := \texttt{discrepancy}_\beta$, based only on the population statistics given in $\texttt{Inputs} \equiv (\{w_g\}, \{\pi_g^y\}, \{\hat{p}_g^y\})$. To illustrate the trade-off between `unfairness` and `error` in $\texttt{disc}_\beta$, Figure 2 shows the solutions for $\beta = 0$ and $\beta = 1$, for a simple problem with two regions of equal weight. In this example, `Inputs` does not preclude perfect fairness.

Suppose first that the classifier is known to be completely fair with respect to the attribute G, according to the equalized odds definition. For instance, a fairness-preserving procedure (e.g., Hardt et al., 2016; Woodworth et al., 2017) could have been used to generate it, or its fairness could have been disclosed to the public. Then for all $y \in \mathcal{Y}, g \in \mathcal{G}$, we have $\alpha_g^y = \alpha_{\text{all}}^y$. Thus, to calculate `error`, it suffices to find the values of $\alpha_{\text{all}}^0, \alpha_{\text{all}}^1$ under this constraint. Eq. (5) and Eq. (4) define the following linear constraints: $\forall g \in \mathcal{G}^+, \alpha_{\text{all}}^1 = r_g + q_g \alpha_{\text{all}}^0$, and $\pi_g^y = 1 \Rightarrow \alpha_{\text{all}}^y = 1 - \hat{p}_g^y$. Thus, `error` can be lower-bounded by minimizing it over $\alpha_{\text{all}}^0, \alpha_{\text{all}}^1 \in [0,1]$ subject to these linear constraints. The pseudo-code for the procedure is given in Appendix A in the supplementary material. Another simple case is when $\beta = 0$. Then, $\texttt{disc}_\beta = \texttt{error}$. This is trivial to lower-bound, since the minimal error under `Inputs` is simply $\sum_g w_g |\pi_g^1 - \hat{p}_g^1|$.

We now proceed to the more challenging task: lower bounding $\texttt{disc}_\beta$ for some $\beta \in (0,1]$ without assuming fairness. Denote $\bar{\alpha} := (\alpha^0, \alpha^1)$. From Eqs. (1), (2), (3):

$$\texttt{disc}_\beta(\{\alpha_g^y\}) = \beta \cdot \min_{\bar{\alpha} \in [0,1]^2} \sum_{g \in \mathcal{G}} w_g \sum_{y \in \mathcal{Y}} \pi_g^y \eta(\alpha^y, \alpha_g^y) + (1-\beta) \cdot \sum_{g \in \mathcal{G}} w_g \sum_{y \in \mathcal{Y}} \pi_g^y \alpha_g^y. \quad (6)$$

Denoting $\tau(a,b) := \beta\eta(a,b) + (1-\beta)b$ and rearranging, we get

$$\texttt{disc}_\beta(\{\alpha_g^y\}) = \min_{\bar{\alpha} \in [0,1]^2} \sum_{g \in \mathcal{G}} w_g \sum_{y \in \mathcal{Y}} \pi_g^y \cdot \tau(\alpha^y, \alpha_g^y).$$

Since $\{\alpha_g^y\}$ are unknown, we cannot calculate $\texttt{disc}_\beta(\{\alpha_g^y\})$ exactly. Instead, we lower-bound this expression under the constraints imposed by `Inputs`.

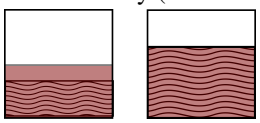
Recall that the free variables are $\{\alpha_g^0\}_{g \in \mathcal{G}^+}$. Since $\alpha_g^y \in [0,1]$ for both $y = 0$ and $y = 1$, it follows from Eq. (5) that the feasible domain of α_g^0 for $g \in \mathcal{G}^+$ is

$$\text{dom}_g := [\max\{-r_g/q_g, 0\}, \min\{(1-r_g)/q_g, 1\}].$$

Input statistics:

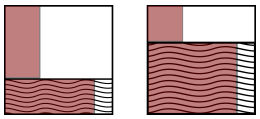
	weight	true pos.	predicted pos.
Left region	1/2	1/3	1/2
Right region	1/2	2/3	2/3

Solution for best accuracy (minimizing $\texttt{disc}_0$):



$\texttt{unfairness} = 1/12, \texttt{error} = 1/12.$

Solution for best fairness (minimizing $\texttt{disc}_1$):



$\texttt{unfairness} = 0, \texttt{error} = 1/4.$

Predicted positives
True positives

Figure 2. Two extreme solutions for a simple 2-region problem

Therefore, given `Inputs`, we can define

$$\text{Obj}(\bar{\alpha}, \{\alpha_g^0\}_{g \in \mathcal{G}^+}) := \sum_{g \in \mathcal{G}^+} w_g \sum_y \pi_g^y \tau(\alpha^y, \alpha_g^y) + \sum_{\substack{g \in \mathcal{G} \setminus \mathcal{G}^+ \\ y: \pi_g^y = 1}} w_g \tau(\alpha^y, 1 - \hat{p}_g^y), \quad (7)$$

and conclude that for any $\{\alpha_g^y\}$ which are consistent with `Inputs`,

$$\texttt{disc}_\beta(\{\alpha_g^y\}) \geq \min_{\substack{\bar{\alpha} \in [0,1]^2 \\ \{\alpha_g^0 \in \text{dom}_g\}_{g \in \mathcal{G}^+}}} \text{Obj}(\bar{\alpha}, \{\alpha_g^0\}_{g \in \mathcal{G}^+}) =: V^*. \quad (8)$$

Our main theorem shows that this high-dimensional optimization problem can be reduced to a small number of one-dimensional problems.

Theorem 4.1. *For $\bar{\alpha} \in [0,1]^2$ and $g \in \mathcal{G}^+$, define the functions*

$$s_g^1(\bar{\alpha}) \equiv s_g^1 := \max\{0, -r_g/q_g\},$$
$$s_g^2(\bar{\alpha}) \equiv s_g^2 := \min\{1, (1-r_g)/q_g\},$$
$$s_g^3(\bar{\alpha}) \equiv s_g^3(\alpha^0) := \alpha^0,$$
$$s_g^4(\bar{\alpha}) \equiv s_g^4(\alpha^1) := (\alpha^1 - r_g)/q_g.$$

Define the set $S_g(\bar{\alpha}) := \{s_g^i(\bar{\alpha})\}_{i \in [4]} \cap \text{dom}_g$, and let

$$\text{Obj}_2(\bar{\alpha}) := \sum_{g \in \mathcal{G}^+} w_g \min_{\substack{\alpha_g^0 \in \\ S_g(\bar{\alpha})}} \sum_y \pi_g^y \tau(\alpha^y, \alpha_g^y) + \sum_{\substack{g \in \mathcal{G} \setminus \mathcal{G}^+ \\ y: \pi_g^y = 1}} w_g \tau(\alpha^y, 1 - \hat{p}_g^y), \quad (9)$$

where $\{\alpha_g^1\}$ are defined according to Eq. (5). Define

$$V_0 := \{0,1\} \cup \{-r_g/q_g, (1-r_g)/q_g\}_{g\in\mathcal{G}^+} \cup \{\hat{p}_g\}_{g:\pi_g^0=1},$$
$$V_1 := \{0,1\} \cup \{r_g, r_g+q_g\}_{g\in\mathcal{G}^+} \cup \{1-\hat{p}_g\}_{g:\pi_g^1=1}. \quad (10)$$

Let V^* as defined in Eq. (8), and define the set

$$\textsf{Sols} := (V_0 \times V_1) \cup \{(v, r_g + q_g v) \mid v \in \mathrm{dom}_g, g \in \mathcal{G}^+\}.$$

Then:

$$V^* = \min_{\bar{\alpha}\in\textsf{Sols}} \mathrm{Obj}_2(\bar{\alpha}).$$

The proof of Theorem 4.1 is provided in Appendix B in the supplementary material. Its main stages are (1) showing that given $\bar{\alpha}$, α_g^y are in a small set and (2) showing that given α^0, α^1 is in a small set and vice versa.

Note that $\mathrm{Obj}_2(\bar{\alpha})$ is easy to calculate for a given $\bar{\alpha}$, since it involves only minimizations on the small finite sets $S_g(\bar{\alpha})$. Moreover, Theorem 4.1 shows that to find V^*, it suffices to minimize $\mathrm{Obj}_2(\bar{\alpha})$ over the set Sols, which includes only $O(|\mathcal{G}|^2)$ solution pairs and $|\mathcal{G}^+|$ one-dimensional solution sets. Thus, a practical procedure for finding V^* can be derived based on Theorem 4.1. For $z \in \mathcal{G}^+$, define $\mathrm{Obj}_3^z(\alpha^0) := \mathrm{Obj}_2(\alpha^0, r_z + q_z\alpha^0)$, and let $V_z = \min_{\alpha^0\in\mathrm{dom}_g} \mathrm{Obj}_3^z(\alpha^0)$. Then, we have $V^* = \min\{\min_{\bar{\alpha}\in V_0\times V_1} \mathrm{Obj}_2(\bar{\alpha}), \min_{z\in\mathcal{G}^+} V_z\}$. The algorithm for finding V^* up to a given tolerance is given in Alg. 1. Alg. 1 solves $O(|\mathcal{G}|)$ one-dimensional minimization problems for Obj_3^z, using the procedure `MinObj3`$(z, \gamma, \beta,$ `Inputs`$)$, which returns a value of α^0 that minimizes Obj_3^z up to a tolerance of γ. This procedure can be implemented, for instance, by bounding the derivative of Obj_3^z and searching on a sufficiently fine grid with respect to the requested tolerance. The time complexity of Alg. 1 is linear in $|\mathcal{G}|$ times the complexity of `MinObj3`.

5. Experiments

We show two types of experiments: First, we test the tightness of the lower bound we obtain for $\mathtt{disc}_\beta$, by comparing it with the true $\mathtt{disc}_\beta$ for the classifier, as calculated using labeled data. Our results suggest that in a large fraction of the cases, the lower bound is within a reasonable factor of the true discrepancy. In the second set of experiments, we demonstrate possible uses and outcomes of the lower-bounding procedure. We study several classifiers for which we only have aggregate statistics, calculate lower-bound Pareto curves of the trade-off between unfairness and error for each classifier, and discuss how these curves can help in decision making. In all the experiments below, we considered classification of US-based individuals, and the attribute of interest was the state of residence (or work) of the individual. Matlab code for Alg. 1, as well as experiment data and code, are available at `https://github.com/sivansabato/bfa`.

Algorithm 1 Finding a lower bound for $\mathtt{discrepancy}_\beta$

Input: $\mathtt{Ins} \equiv (\{w_g\}_{g\in\mathcal{G}}, \{(\pi_g^y, \hat{p}_g^y)\}_{g\in\mathcal{G}, y\in\mathcal{Y}})$, $\beta \in [0,1]$, tolerance $\gamma > 0$

Output: A value $V \in [V^*, V^* + \gamma]$, where V^* (see Eq. (8)) is the $\mathtt{disc}_\beta$ lower-bound; The values of `unfairness` and `error` that obtain V.

For $g \in \mathcal{G}^+$, $r_g \leftarrow 1 - \hat{p}_g^1/\pi_g^1$ and $q_g \leftarrow 1/\pi_g^1 - 1$.
Set V_0, V_1 as in Eq. (10).
for $z \in \mathcal{G}^+$ **do**
    $\alpha_z^0 \leftarrow$ `MinObj3`$(z, \gamma, \beta,$ `Ins`$)$; $V_z = \mathrm{Obj}_3^z(\alpha_z^0)$.
end for
$\mathrm{Pairs} \leftarrow (V_0 \times V_1) \cup \{(\alpha_z^0, r_z + q_z\alpha_z^0) \mid z \in \mathcal{G}^+\}$.
$\hat{\alpha} \leftarrow \mathrm{argmin}_{\bar{\alpha}\in\mathrm{Pairs}} \mathrm{Obj}_2(\bar{\alpha})$; $V \leftarrow \mathrm{Obj}_2(\hat{\alpha})$.
$\forall g \in \mathcal{G}^+, \hat{\alpha}_g^0 \leftarrow \mathrm{argmin}_{\alpha_g^0\in S_g(\hat{\alpha})} \sum_y \pi_g^y \tau(\hat{\alpha}^y, \alpha_g^y)$; $\hat{\alpha}_g^1 \leftarrow r_z + q_z\hat{\alpha}_g^0$.
$\forall g \in \mathcal{G} \setminus \mathcal{G}^+$ and y s.t. $\pi_g^y = 1$, set $\hat{\alpha}_g^y \leftarrow 1 - \hat{p}_g^y$.
`unfairness` $\leftarrow \sum_{g\in\mathcal{G}} w_g \sum_y \pi_g^y \tau(\hat{\alpha}^y, \hat{\alpha}_g^y)$.
`error` $\leftarrow \sum_{g\in\mathcal{G}} w_g \sum_y \pi_g^y \hat{\alpha}_g^y$.
Return V, `unfairness`, `error`.

5.1. Tightness of the Lower Bound on `discrepancy`

In the first experiment, we used the UC Census (1990) data set (Dua & Graff, 2019) to generate hundreds of classifiers, on which we could test the tightness of the lower bound. The US Census data set has ≈ 2.5 Million records, with 124 attributes for each record. We used for the experiment the ≈ 1.1 Million records in which the "state of work" attribute was present. We split this data into two halves at random, using one half as a training set to generate classifiers, and the other half as a test set to calculate the aggregate statistics of the classifier, as well as its true `unfairness` and `error`. For each attribute in the data set other than the state of work, and for each value v of the attribute, we generated a classification problem in which the examples are the records without this attribute and without the "state of work" attribute, and the label of a record was 1 if the attribute had value v. For attributes with more than 10 values, we binned the values of the attribute to 10 bins, and the label was 1 if the attribute had value at least v. If the resulting classification problem had more than 99% of the examples assigned to the same label, this classification problem was discarded. This process resulted in 410 classification problems. For each classification problem, we generated a classifier using linear regression with standard a Matlab package.

We ran Alg. 1 on `Inputs`, as calculated for each of these classifiers on the test set. The fraction of the population in each state, $\{w_g\}$, was also calculated based on the test set. First, we used Alg. 1 to calculate a lower bound on $\mathtt{disc}_1 \equiv$ `unfairness`. We then calculated the ratio between the true `unfairness` (as calculated on the same

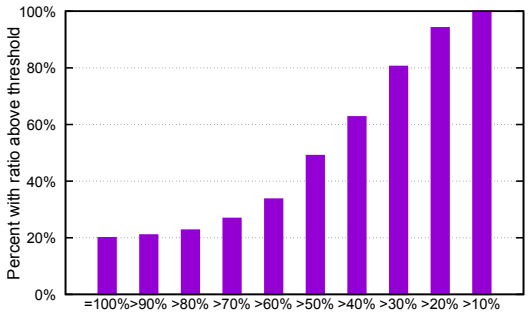


Figure 3. `US Census` experiments. For each threshold, the fraction of classifiers for which the ratio between the `unfairness` lower bound returned by Alg. 1 and the true `unfairness` was more than this threshold.

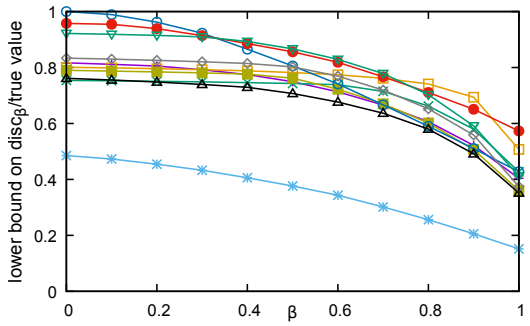


Figure 4. `US Census` experiments. For 10 randomly selected classifiers, the ratio between $\texttt{disc}_\beta$ lower bound and the true $\texttt{disc}_\beta$, as a function of β. 1 is optimal.

test set) and the lower bound. Figure 3 shows the fraction of the classifiers which achieved a ratio above a given threshold. The median ratio was 48%. All classifiers obtained a ratio of at least 11%. Next, we used Alg. 1 with a range of values of β, for 10 randomly selected classifiers. Figure 4 plots the ratio between the true $\texttt{disc}_\beta$ and our obtained lower bound for $\beta \in [0, 1]$. Here too, it can be observed that most lower bounds are reasonably tight.

5.2. Pareto Curves for Classifiers with Unknown Rates

In the next set of experiments, we study several classifiers for which we only have aggregate statistics, and produce Pareto curves of the resulting lower-bounds, demonstrating the (best possible) trade-off between `unfairness` and `error`. The population fraction in each state, w_g, was obtained from official records (US Census Bureau, 2019). We generated Pareto curves for the studied classifiers by running Alg. 1 for $\beta \in \{0, 0.01, \ldots, 0.99, 1\}$, and extracting the pairs of $(\texttt{unfairness}, \texttt{error})$ that minimized $\texttt{disc}_\beta$ for each of the values of β.

In the first experiment, we consider a classification problem of identifying people diagnosed with a certain type of cancer from their search queries via the Bing search engine. The end goal was to develop an anonymous patient cohort, as discussed, e.g., in Soldaini & Yom-Tov (2017). We studied the 18 cancer types listed in CDC & NCI., 2019, in which the true-positive rates (incidence rates) $\{\pi_g^1\}$ in each state are reported. For each cancer type, we constructed a classifier that predicted the label for US-based search-engine users. The label was predicted positive if the user mentioned the cancer in queries between January 1st and June 30th, 2019. We do not have individual validation data connecting users to their true diagnostic status. For each classifier, we calculated the rate of predicted positives $\{\hat{p}_g^1\}$ in each US state, rescaling by the fraction of search-engine users in that state.

For each classifier, we wish to discover a best-case accuracy-fairness trade-off, so as to identify classifiers that might be useful for a future, more detailed study. We note that unfairness may ensue using this classification method, for instance due to differences in health literacy between states. Out of the 18 cancer types, we studied the 10 types for which the ratio between the overall positive prediction and true positive rate was within $[\frac{1}{2}, 2]$. The Pareto curves of these classifiers are reported in Figure 5. To allow easy comparison between cancer types, the the values for each classifier are normalized in the plot by the overall true-positive rate of that cancer type. Thus, values close to 1 indicate a poor classifier. Recall that we do not have the ground truth for these classifiers, thus we cannot compare to it. However, since we proved that Alg. 1 gives a lower bound, the true $(\texttt{unfairness}, \texttt{error})$ pair of each classifier is necessarily above its Pareto curve. We conclude from the graph that the classifier for lung cancer is the most promising for a future study, while many of the other classifiers necessarily perform poorly.

In the next experiment, we study pre-election polls, and use them to provide discrepancy lower bounds on the classifiers they might represent. We obtained data from 10 pre-election polls of the 2016 US Presidential elections (Five Thirty Eight, 2016). The label was set to positive if the individual voted for the Democratic candidate. Each poll predicted a voting rate for the candidate in each state. The true positive rate was obtained from the actual results of the presidential elections in each state (Federal Elections Commission, 2016). Treating each poll's voting prediction as the aggregate statistics of an unknown classifier, we calculated the Pareto curves representing the best-possible combinations of `unfairness` and `error` for these classifiers. The resulting Pareto curves, shown in Figure 6 by date of poll publication, can be used to compare the classifiers that could be underlying these polls. For instance, although some of the more accurate polls are also the latest ones, this is not always the case. Moreover, some pairs of

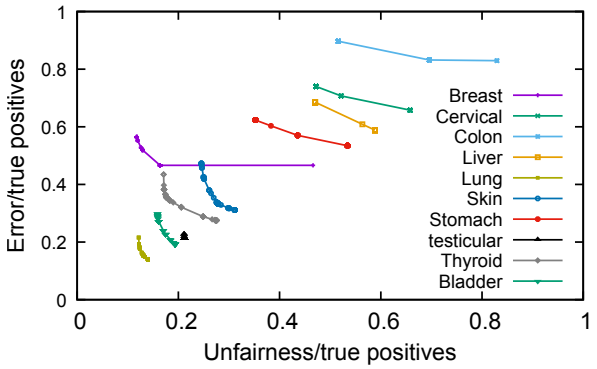


Figure 5. Pareto curves of (`unfairness`,`error`) for each classifier (cancer type). Values are normalized by the overall true-positive rate of the relevant cancer type.

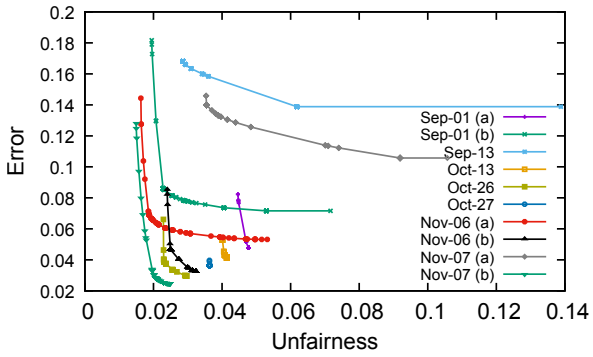


Figure 6. Pareto curves for the classifiers induced by pre-election polls from the 2016 US Presidential elections. The key indicates the date in which the poll was published.

polls are incomparable, since they perform better in different parts of the curve. For instance, the Nov-06(b) poll has a lower error bound than the Nov-06(a) poll if its unfairness is larger than approx. 2%, but it cannot have unfairness smaller than 2%, while the other poll can, hinting perhaps at a more biased polling methodology. This information can be used to further analyze the polling and prediction strategies employed in the various polls.

In the last experiment, we use the proposed method to explore variation in cancer mortality rates across US states. Rates of cancer diagnosis and mortality in each US state, for 10 cancer types, were taken from the data published in CDC & NCI., 2019.[1] The true-positive rates $\{\pi_g^1\}$ for each cancer type were set to the fraction of people in each state who died from the given cancer. We generated aggregate statistics for a simulated classifier, which predicts mortality of a diagnosed individual in each state with a probability

[1]In cases where data from some states was missing, we removed these states from the list of regions and renormalized the weights of the other states.

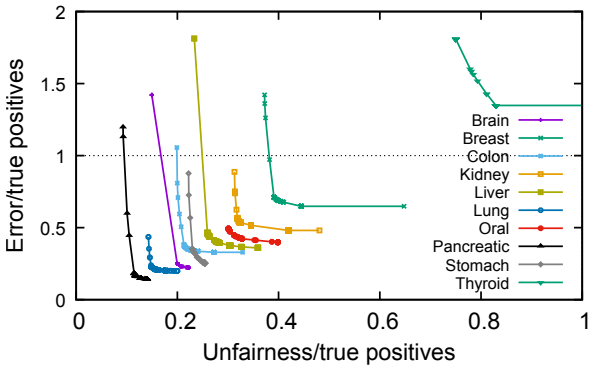


Figure 7. Pareto curves based on cancer diagnosis and mortality rates in each state. Values are normalized by the overall true-positive rate of the relevant cancer type.

equal to the *overall* mortality rate, which is the ratio between the overall mortality and the overall diagnosis rate. Thus, our simulated classifier is based on the premise that in each state, the mortality rate is the same. Any deviation from this premise would result in a classifier which cannot be fully accurate or fair. In this context, a high `unfairness` value may be interpreted as a large fraction of individuals who have a non-typical variation of the disease, while a high `error` may indicate a large fraction of the population whose mortality differs from the expected mortality, perhaps due to a difference in access to health services. The Pareto curves of the classifiers are shown in Figure 7; values for each cancer type are rescaled by the rate of true positives of that cancer. It can be observed, for instance, that in several of the cancer types, allowing a small amount of `unfairness`, interpreted as modeling a fraction of the population as having non-typical variations of the disease, leads to a significantly smaller error, interpreted as a small difference in mortality rate between the states, on the population with the typical variation of the disease.

6. Discussion

In this work, we showed that useful bounds on fairness and accuracy can be provided for classifiers based only on the aggregate statistics on predicted positive rates and true positive rates. We defined a new `unfairness` measure to facilitate the study of classifiers that are not completely fair, and provided an efficient and practical procedure which provably lower-bounds a given trade-off between fairness and accuracy. In future work, we plan to generalize the methodology suggested here to other group-fairness definitions, and to more general supervised learning schemes such as multiclass classification and regression. Our experiments show how this procedure allows tackling problems in social and health studies, as well as in classifier design.

A natural question is whether meaningful *upper* bounds on $\texttt{discrepancy}_\beta$ can also be similarly obtained. However, even if the classifier has zero error, upper bounding the error must take into account the possibility of no overlap between the positive predictions and the true positives, thus it would be at least twice the positive rate. Therefore, obtaining a meaningful upper bound on the discrepancy is a challenging open problem that may require additional modeling assumptions.

References

Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. *NIPS Tutorial*, 2017.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Black, E., Yeom, S., and Fredrikson, M. Fliptest: Fairness auditing via optimal transport. *CoRR*, abs/1906.09218, 2019.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3992–4001. Curran Associates, Inc., 2017a.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3992–4001. Curran Associates, Inc., 2017b.

CDC and NCI. United states cancer statistics: Data visualizations, 2019. URL https://gis.cdc.gov/Cancer/USCS/DataViz.html.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348. ACM, 2019.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2791–2801. Curran Associates, Inc., 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2019. URL http://archive.ics.uci.edu/ml.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Federal Elections Commission. Federal elections 2016, 2016. URL https://transition.fec.gov/pubrec/fe2016/federalelections2016.pdf.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Five Thirty Eight. National presidential polls, november 8th, 2016, 2016. URL https://projects.fivethirtyeight.com/2016-election-forecast/national-polls/.

Goel, N., Yaghini, M., and Faltings, B. Non-discriminatory machine learning through convex fairness criteria. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. Satisfying real-world goals with dataset constraints. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2415–2423. Curran Associates, Inc., 2016.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, pp. 2, 2016.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Heidari, H., Ferrari, C., Gummadi, K., and Krause, A. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In Bengio, S., Wallach, H.,

Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1265–1276. Curran Associates, Inc., 2018.

Jackson, C., Best, N., and Richardson, S. Improving ecological inference using individual-level data. *Statistics in medicine*, 25(12):2136–2159, 2006.

Jackson, C., Best, N., and Richardson, S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1):159–178, 2008.

Johndrow, J. E. and Lum, K. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4066–4076. Curran Associates, Inc., 2017.

McDuff, D., Ma, S., Song, Y., and Kapoor, A. Characterizing bias in classifiers using generative models. In *Advances in Neural Information Processing Systems 32*, pp. 5404–5415. Curran Associates, Inc., 2019.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118, 2018.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Soldaini, L. and Yom-Tov, E. Inferring individual attributes from search engine queries and auxiliary information. In *Proceedings of the 26th international conference on World Wide Web*, pp. 293–301, 2017.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2239–2248, 2018.

Sun, T., Sheldon, D., and OConnor, B. A probabilistic approach for learning with label proportions applied to the us presidential election. In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 445–454, Nov 2017. doi: 10.1109/ICDM.2017.54.

US Census Bureau. Annual estimates of the resident population for the united states, regions, states, and puerto rico: April 1, 2010 to july 1, 2019, 2019. URL `https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/totals/nst-est2019-01.xlsx`.

Verma, S. and Rubin, J. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7. IEEE, 2018.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1920–1953, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

Wu, Y., Zhang, L., and Wu, X. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, pp. 3356–3362. ACM, 2019.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017a.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970, 2017b.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28(3) of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.